Performance Comparison of Machine Learning Algorithms for the Indoor Positioning System

W.A.H. Gimhani, R.S. Madara, A. K. D. D. Thilakarathna , M.R. Udawalpola
Department of Electrical and Information Engineering
Faculty of Engineering
University of Ruhuna.
Galle, Sri Lanka

Abstract — In this paper the problem of indoor positioning based on readings from its embedded sensors, utilizing machine learning methodologies is investigated. It is useful to differentiate algorithms based on computational performance rather than classification accuracy alone. As although classification accuracy between the algorithms is similar, computational performance can differ significantly and it can affect to the quality of the indoor positioning device if it takes a considerable time to display the current position [1]. So the objective of this paper is to perform a comparative analysis of 8 machine learning algorithms namely, K means, Hierarchical, Gaussian mixture , density based , Nearest neighbor, feature agglomeration , Linear Regression and Logistic Regression. A real world indoor fixed environment is considered a large dataset of 1278 data points is collected. Then the performance of the above mentioned machine learning algorithms are examined. In this paper the processing time of the different machine learning techniques are being estimated by considering the collected data set, over a fixed indoor area of 52.545 m^2. The paper is organized as follows. In Section I, introduction and background analysis of the research is included and in section II, other indoor localization systems and their limitations are examined. In Section III, our application and data collection Process, the testing environment, and the Methodology of our analysis are being described briefly. Section IV comprises the results of our analysis. Finally, the paper concludes with a discussion of future directions for research by eliminating the problems existing with the current research methodology.

Keywords- Machine Learning, Sensor Fusion, Kalman Filter

I. INTRODUCTION

Nowadays, developing Indoor Positioning System has become an attractive research topic due to the increasing demands on Location-Based Service in indoor environments [9]. A positioning system tries to determine the location of an object in space [3]. One such positioning system is the coordinate system with 2 or 3 dimensions. There are different types of the coordinate system named the Cartesian coordinate system, Polar coordinate system and Spherical coordinate system. In the three dimensional Cartesian coordinate system, the axes are determined as *X, Y* and *Z*. A point *P* in this system has *X, Y* and *Z* coordinates represented as P=(X, Y, Z). The reference point *P* can be represented as (0,0,0) [6].

There are several techniques have been found today for the accurate position estimation. The first one is the Sensor Fusion Technique, combining sensory data from two or more types of sensors to update the state of a system. Sensor fusion uses Kalman Filters that combine the values from the several sensors to give the final accurate values [2], [4], [8]. The other techniques are machine learning based techniques. Machine learning is a field of artificial intelligence dealing with algorithms that improve performance over time with experience. Supervised learning algorithms for regression are trained on data with the correct value given along with each variable This allows the learner to build a model, based on the attributes that is best fit the correct value. By giving more data to the algorithm the model can be improved [10].

Learning can be described in this way as improving performance. The measure of performance is how well the algorithm predicts the regression value given a set of variables or attributes [5]. Machine learning algorithms provide excellent solutions for building models that generalize well given large amounts of data with many attributes by discovering patterns and trends in the data. Machine learning algorithms are a natural solution for sifting through these large datasets and determining the important pieces of information for localization, building accurate models to predict an indoor position [6].

Machine learning algorithms may also provide a fast and efficient method for indoor tracking, which will be often more useful to applications than static localization [6].
In this study, a performance analysis of a wide range of machine learning algorithms using real-world data for localization is performed.

II. PROBLEM STATEMENT AND RELATED WORK

Machine learning algorithms are an advanced and efficient solution for determining the accurate models to predict an indoor position. [6] But the most suitable machine learning algorithm with maximum performance have to be decided, as our intended purpose is to increase the accuracy and performance of the indoor positioning system than the existing positioning systems. Since in an indoor positioning and navigation system if the processing time of algorithm is high, the position estimation can become inaccurate due to the user's movement during calculation of the position. This problem can be overcome by implementing these kind of research works.

Indoor localization research has garnered a good deal of interest from both academia and industry, with numerous systems being proposed using a variety of technologies. A major disadvantage of many of these systems such as infrared and ultrasound is that they require dedicated sensors, substantial infrastructure changes and as a result, incur a significant cost to deploy [8].
Much of the existing research focuses on the achievable accuracy (classification accuracy) of different machine learning algorithms. The studies have shown that a number of different algorithms are able to achieve high classification accuracy. The effect of using different sets of statistical features on the same dataset has seen little investigation [5].

Effort has been made to devise localization systems that require little to no infrastructure change using Bluetooth, and Wi-Fi signal strengths with some success. The systems developed using Wi-Fi signal strengths for localization show promise but have yet to receive widespread adoption [6]. These systems can be divided into two categories. Those using a fingerprinting approach using algorithms for "nearest neighbor in signal space" and those using more complex signal propagation algorithms to determine a device's distance from the access points in range [6].

Localization systems that use a nearest neighbor in signal space approach, require a collection of data points throughout the room or building [5]. To predict a position, a new set of attributes constituting a new data point is compared with every point in the classified dataset. Depending on the implementation of the k-Nearest Neighbor algorithm, the coordinates of the closest point are used as the coordinates for the new point or an average of k closest points can be used with different weights [5]. These instance-based machine learning approaches can achieve accuracies up to 2 meters on average, but current research is limited in that only one or a few algorithms are considered and has not taken into account many of the sensors available in most modern mobile devices[5].

Further, these algorithms are limited by the size of the dataset since a large dataset will require a substantial amount of time to predict a position, hindering real-world deployment. Systems that build signal propagation maps for a building have achieved similar accuracy. But generally require a great deal of information about the Wi-Fi access points that may not be known or readily available. Additionally, these models take only Wi-Fi signal strengths into account for localization. Hence these may not be as accurate as models that account more variables [5].

Lim et al. conducted an extensive survey of 33 algorithms across 32 diverse datasets. They found that the algorithms showed similar classification accuracy. But they showed quite different training performance times. (For a given dataset and complementary features). Hence they recommended that users should select algorithms based on criteria, such as model interpretability or training time [10].

III. METHODOLOGY

The localization process consists of two distinct phases. They are data collection and analysis. Firstly an indoor area which is subjected to less disturbances from humans and consist of less number of equipments was selected and the points are marked.

In an initial scan of the various signals received throughout the building, a few portable hotspots likely from people in the building tethering their devices were detected. As these signals would not remain constant, they were not included as attributes for the algorithms to train on. Only the Wi-Fi signals that were part of the building infrastructure were stored.

A. Data Acquisition

The readings from the three sensors (magnetometer, accelometer and gyroscope) are combined together by using sensor fusion in such a way that only three outputs are obtained (KalmanX, KalmanY, Kalman Z) from all the three sensors.
The device equipped with the sensors such as magnetometer (HMC5883), accelometer and gyroscope (MPU 6050) and Arduino Wi-Fi shield (CC3000).

The data collection phase consists of placing the device above the points which were marked previously throughout the selected environment and obtaining the readings of the Wi-Fi signal strengths and data from the other sensors in the device. The sensor values which are associated with a user-provided location and the actual coordinates of those positions were included to an Excel sheet manually at the same time. In our case, a single data point, consists of 4 attributes corresponding to values from each sensor on the device and they were obtained 0.95m above from the ground level.

In total, 1278 data points were considered in the 2nd floor of the Department of Electrical and Information Engineering, with an area of 52.545 m^2. Measurements were taken in a grid based on the floor in the building and all measurements are relative to the reference point. 0.2m distance have between every two consecutive points, meaning if the reference point is considered as (0,0) the point at coordinate (20cm , 40cm), 0.2m away from X direction (front from origin) and 0.4m away from Y direction (left side from origin).

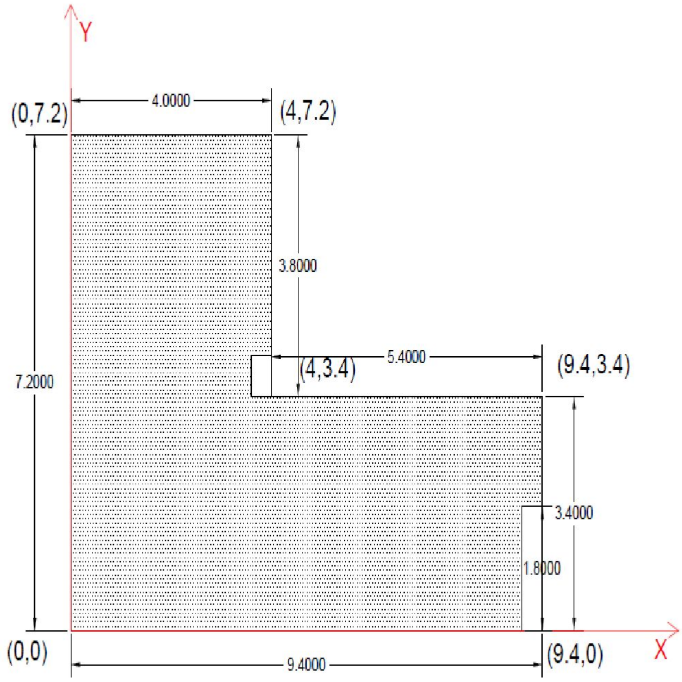

Fig.01. The map of Collected data in indoor environment

The map of collected data points can be seen in Figure 01 and all the dimensions are marked in meters.
The missing data points which have left blank in figure 01, are caused by obstacles in the building such as pillars, cupboards and bag racks which made it impossible to record data at those points.

Collection of data involved resetting of the device at a new point, initiating a scan of Wi-Fi network strengths in range of the device and displayed all those sensor values and Wi-Fi signal strengths on the remotely connected computer through a 5m length data cable. This whole process took an average time of 10.73 seconds to complete.

B. Analysis

After the data collection process, the Excel file (which is comprised of collected data) was converted to ".csv" file format and was used that ".csv" file as an input to the python programme by using the "pandas" library. Then the all the sensor values were stored as variable arrays by using "numpy" library. The machine learning algorithms were applied to those stored data sets and the data acquisition time of each and every algorithm was obtained. The data set was divided into several sets and obtained the execution time and a graph was obtained for the easiness of comparison.

Here Clustering Algorithms (K-means, Hierarchical Algorithm, Gaussian Mixture Algorithm, Density based Algorithm) Supervised learning Algorithms (Linear regression , logistic regression , nearest neighbour) , feature agglomeration have considered in our execution time comparison.

IV. RESULTS

For performing comparative analysis, this paper principally focuses on the time taken to form clusters (in clustering algorithms), and number of iterations.

Result shows that Density based algorithm takes lowest time i.e. 0.008 seconds for our 1200 number of data sets.

Hence in terms of efficiency, density based clustering algorithm produce better result as compared to other algorithms and the K mean algorithm is the second best algorithm in order to use for an indoor positioning system. In the table below are the execution times for 1200 number of data.

TABLE 1: THE TABLE OF EXECUTION TIME FOR DIFFERENT ALGORITHMS

Name of the Algorithm	Execution Time of the Algorithm (s)
K mean	0.019
Hierarchical Algorithm	0.036
Gaussian Mixture	0.040
Density based Algorithm	0.008
Nearest neighbor	0.096
Feature Agglomeration	0.059
Linear Regression	0.031
Logistic Regression	0.049

For the execution time values obtained for the six data sets (200, 400, 600, 800, 1000, 1200), the following graph in below Figure 02 can be obtained.

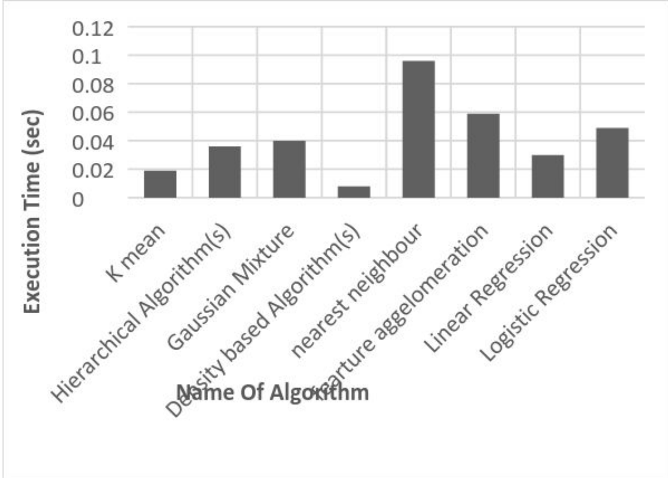

Fig. 2. The Graph of Execution time Vs name of algorithm

V. CONCLUSION AND FUTURE WORKS

Much of this existing research focuses on the achievable accuracy (classification accuracy) of different machine learning algorithms [11].

These experiments have used different (thus not comparable) datasets and features. The process of defining appropriate features, performing feature selection and the influence of this on classification and computation performance has not been studied.

So in this work, the execution time of a several number of machine learning algorithms which are used widely in indoor localization systems have been examined. The processing time of an algorithm is a critical point to be considered. Because it can increase the error in an indoor positioning and navigation system, due to the time taken to calculate a position [11]. In the case of Gaussian Mixture algorithms, this may be due to the time taken to calculate a position. Because Gaussian mixture algorithm takes an average of 0.04 seconds to calculate a position and the user is moving during the calculation. So predicted position may be indicative of the user’s position 0.04 seconds in the past by the time computation finishes.

As this paper represents a preliminary investigation, there are a number of potential avenues for further work, such as an in-depth evaluation of as classification accuracy of different machine learning algorithms. In a wider context, investigating the robustness of ML classification (for instance training on a data from one location and classifying data from other locations) and a comparison between ML and non-ML techniques on an identical dataset would also be valuable [10].

REFERENCES

[1] Indoor Positioning using Sensor-fusion in Android Devices,
School of Health and Society Department of Computer Science Embedded Systems by Authors Ubejd Shala Angel Rodriguez

[2] Kalman Filter based Indoor Mobile Robot Navigation Md Anam Mahmud, Md Sayedul Aman, Haowen Jiang, Ahmed Abdelgawad, Kumar Yelamarthi School of Engineering and Technology Central Michigan University Mount Pleasant, MI, 48859, USA

[3] Mobile Robot Positioning & Sensors and Techniques by J. Borenstein1 , H.R. Everett2 , L. Feng3 , and D. Wehe4

[4] Multisensor Funtion For Mobile Robot Positining. by Agostino Martinelli ,Francesco Martinelli , Salvatore Nicosia,Paolo Valigi

[5] Machine Learning for Indoor Localization Using Mobile Phone-Based Sensors ,A thesis submitted by David Mascharka and Eric Manley Department of Mathematics and Computer Science Drake University Des Moines, Iowa

[6] Smart indoor localization using machine learning techniques, A thesis Submitted by Viney Anand Ugave Department of Electrical and Computer Engineering In partial fulfillment of the requirements For the Degree of Master of Science Colorado State University

[7] A paper Published in Innovations in Intelligent Systems and Applications (INISTA), 2015 International Symposium on Date 2-4 Sept. 2015.

[8] Multi sensor data fusion, a thesis submitted by P. Grossmann to the GEC journal of Technology, in 1998.

[9] Pedestrian localization for indoor environments A dissertation submitted for the degree of Doctor of Philosophy Oliver J. Woodman St Catharines College September 9, 2010.

[10] A Preliminary Performance Comparison of Five Machine
Learning Algorithms for Practical IP Traffic Flow Classification a research paper submitted to the Centre for Advanced Internet Architectures (CAIA) Swinburne University of Technology, Melbourne, Australia by Nigel Williams, Sebastian Zander, Grenville Armitage

[11] A Comparative Analysis of Clustering Algorithms, A Research paper by Raja bala , Sunil Sikka ,JUhi singh from Amity university , India.